Semantic Web journal 0 (2015) 1–0
IOS Press

EventMedia: a LOD Dataset of Events Illustrated with Media

Editor(s): Pascal Hitzler, Wright State University, Dayton, OH, USA
Solicited review(s): Christophe Guéret, Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences, Den Haag, The Netherlands; Erik Wilde, School of Information, UC Berkeley, USA; Amit Joshi, Wright State University, Dayton, OH, USA

Houda Khrouf and Raphaël Troncy
EURECOM, Multimedia Department, Campus SophiaTech, France
E-mail: {khrouf,troncy}@eurecom.fr

Abstract. An ever increasing amount of event-centric knowledge is spread over multiple web sites, either materialized as calendar of past and upcoming events or illustrated by cross-media items. This opens an opportunity to create an infrastructure unifying event-centric information derived from event directories, media platforms and social networks. In order to create such infrastructure, EventMedia relies on Semantic Web technologies that ensures seamless aggregation and integration of disparate data sources, some of which overlap in their coverage. In this paper, we present the EventMedia dataset composed of events descriptions associated with media and interlinked with the Linked Open Data cloud. We describe how data has been extracted, converted, interlinked and published following the best practices of the Semantic Web.

Keywords: Events, Linked Data, Social Media, LODE Ontology

1. Introduction

In their daily life, people naturally organize their personal data according to occurring events: holiday, wedding, birthday party, concert, etc. Events are indeed a natural way for referring to any observable occurrence grouping persons, places, times and activities [1]. Events are also observable experiences that are often documented by people through different media. Nowadays, social services host a large amount of information about events, illustrative media and social connections between participants. However, this information is often spread and locked in amongst these services providing limited event coverage and no interoperability of the description [2]. Aggregating these heterogeneous sources into one unified platform is the aim of the EventMedia project leveraging on the benefits of Semantic Web technologies.

One vision of the Web of Data is to organize the data silos in a structured way which can be understood by machines and easily exploited by humans. This requires the use of common vocabularies for the integration of fragmentary information into a logically coherent knowledge. To achieve this vision, a growing number of RDF datasets have been published in the Web of Data covering diverse domains such as digital libraries, government, health, media or more generally encyclopedic data. In this work, our goal is to introduce an event-domain RDF dataset and to investigate the underlying connections between event centric data on the Web. While wishing to create such dataset, we are aware that event web directories already exist such as Last.fm[1], Eventful[2], Upcoming[3] and Faceook[4] to name a few. However, these services provide limited coverage of events and insufficient browsing options for decision support (e.g. lack of location map and media). As a solution, there is a need to create an infras-

[1] http://www.last.fm
[2] http://www.eventful.com
[3] http://www.upcoming.org
[4] http://www.facebook.com

tructure that enhances data exploration with the flexibility and depth afforded by Semantic Web technologies, and allows users to discover meaningful relationships amongst events. Therefore, we have created the EventMedia dataset which is obtained from four large public event directories namely Last.fm, Eventful, Upcoming and Laynrd[5], and from two large media directories namely, Flickr and Twitter. Our strategy is to select popular sources, but this list is non-exhaustive.

The remainder of this paper is structured as follows. We explain how the data is collected (Section 2) and converted into an RDF model (Section 3). We present an overview of the EventMedia dataset in Section 4, and we describe how we interlinked it with the other LOD datasets in Section 5. Then, we describe two web application in Section 6, and we outline the future work in Section 7.

2. Crawling and Aggregating Data

In this section, we describe how the data has been collected and interlinked either statically using a REST-based crawler or dynamically using a live extractor.

2.1. REST-based data Crawler

Crawling data from multiple services is in general a time consuming task due to the lack of haromonization in different specifications and policies of REST APIs (Application Programming Interface). This imposes a need to create tools providing a seamless and flexible way to crawl data from multiple services. Such tools should be able to address many tasks such as policy management, requests chaining, data integration or merging response schemas. We propose a framework that supports those tasks and unifies information into a meaningful data model. This framework is composed of two main components: the Unified REST Module and the Scraping Processor as illustrated in Figure 1. The first module is based on a RESTful service that allows for the unification of various Web APIs by exploiting their commonality in terms of described methods, inputs and responses. Each source API (e.g. Eventful API) is associated with a descriptor file which represents the API parameters such as root URL, API key, and a set of query objects. Then, each query object represents a mapping between our REST URL pattern and the source API URL pattern which describes a REST method and its input parameters. In order to manage the request chaining, we define two types of query objects: (i) the first type is related to first-order methods used to search for the main objects such as events and media, (ii) the second type is related to other methods used to fetch the descriptions of secondary objects such as artists, locations, attendees, etc. Overall, we have created three REST methods to search for events, photos and videos, respectively. These methods have as input a set of parameters such as the original sources (e.g. last.fm, eventful, etc.) and other additional filters (e.g. category, location, date, etc.). Thus, the user can request in parallel multiple Web services by specifying the list of sources into one request.

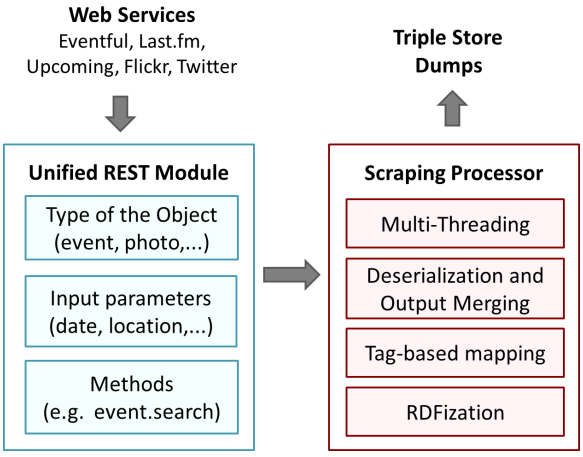

Fig. 1. The Rest-based Crawler Architecture

Besides the RESTful service, the Scraping Processor manages four important tasks. The first task enables multi-threading to reduce the amount of time usually required to query multiple web services. The remaining tasks deal with data processing, starting from JSON de-serialization to RDF conversion and loading into a triple store. More precisely, data retrieved is de-serialized and exported into a common schema providing descriptions of a set of objects, namely; event, location, agent, user, photo and video. Then, we use a tag-based mapping by consuming some metadata, not only to establish links between events and media, but also to enrich their descriptions with additional information from external datasets. This framework is meant to ease the addition of new APIs used to collect events and media. It also offers other REST methods to track or stop the scraping processes.

[5] http://www.lanyrd.com

Finally, a web dashboard has been developed in order to offer graphical functionalities that help monitor the scraping task. It provides practical widgets to help build a query by filtering some parameters and track the scraping process. It is available online at `http://eventmedia.eurecom.fr/dashboard`.

2.2. Tag-based Mapping

A recent user-centric study [2] highlights the importance of media to provide visual information which support decision making. This study motivated us to enrich event views with media by exploring the overlap in metadata between four popular web sites, namely Flickr as a hosting web site for photos and videos, and Last.fm, Eventful and Upcoming as a rich documentation of past and upcoming events. Note that explicit relationships between events and photos exist using machine tags such as `lastfm:event=XXX`. We have been able to convert the descriptions of more than 1.7 million photos which are indexed by nearly 140.000 events. We further leverage the machine tags to create links between various directories such as `foursquare:venue=XXX` used to link venues descriptions with Foursquare[6] directory (a location-based service), and `musicbrainz:artist=MBID` used to link artists descriptions with MusicBrainz[7] directory (an open music database). Similarly, we also exploit the existing overlap between Twitter and Lanyrd (a social conference directory) where each conference is associated with a Twitter hashtag. Thus, we have been able to convert the descriptions of more than 530.000 tweets which are indexed by nearly 1.167 conferences.

2.3. Live Data Extraction

New events are taking place everyday and people keep sharing an ever-growing amount of media. Such evolution requires a real-time processing that retrieves fresh data and updates the triple store. To achieve this, we developed a live extractor which consumes the feeds provided by some Web services. More precisely, we use the Flickr feeds[8] which contains the tag "`*:event=`". Then, a scheduled process reads the feeds every 10 minutes and trigger accordingly the scraping requests to retrieve the descriptions of events and photos. On an average week, we observe 1500 new photos and 130 new events which are added to EventMedia. Similarly, we also use the Lanyrd feeds[9] which provides fresh information about conference including the main hashtag required to retrieve related tweets.

3. RDF Modeling

In this section, we describe our approach to generate RDF triples describing events and media using a variety of existing vocabularies such as the LODE ontology and Media Resources ontology.

3.1. The LODE Ontology

The LODE ontology[10] is a minimal model that encapsulates the most useful properties for describing events. LODE is not yet another "event" ontology *per se*. It has been designed as an *interlingua* model that solves an interoperability problem by providing a set of axioms expressing mappings between existing event ontologies. Hence, the ontology contains numerous OWL axioms stating classes and properties equivalence between models such as MO [3], CIDOC-CRM [4] and DOLCE to name a few. In addition, LODE can be enhanced with mappings to other vocabularies such as Schema.org and DBpedia. Overall, the goal of LODE is to enable an interoperable modeling of the "factual" aspects of events, where these can be characterized in terms of the four Ws: What happened, Where did it happen, When did it happen, and Who was involved. "Factual" relations within and among events are intended to represent intersubjective "consensus reality" and thus are not necessarily associated with a particular perspective or interpretation. We use the LODE ontology together with properties from FOAF, Dublin Core and VCard. Our strategy is to separate events from their interpretations with an emphasis on factual aspects, a design approach that has not been considered in other event models [1]. Figure 2 depicts the metadata attached to the event identified by `3163952` on Last.fm according to the LODE ontology. More precisely, it indicates that an event of type `Concert` has been given on the `21th of May 2012 at 12:45 PM` in the `The Paramount Theatre` featuring the `Snow`

[6] `https://foursquare.com`
[7] `http://musicbrainz.org`
[8] `http://api.flickr.com/services/feeds/photos_public.gne?tags=*:event`
[9] `http://api.lanyrd.com/conferences`
[10] `http://linkedevents.org/ontology/`

Patrol rock band, and one of attendees is the Last.fm user earthcapricor. Using the machine tag of related media, an *owl:sameAs* link is discovered between this event and a similar one announced on Upcoming.

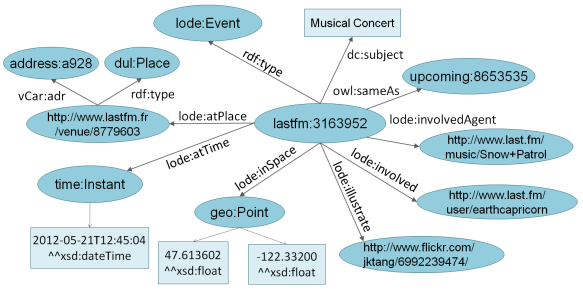

Fig. 2. The *Snow Patrol Concert* described with LODE ontology

3.2. Media modeling

To describe media, we re-use two popular vocabularies, namely: the W3C Ontology for Media Resources[11] to represent photos and videos, and SIOC[12] to represent tweets, status, posts and slides. The Ontology for Media Resource is a core vocabulary which covers basic metadata properties to describe media resources. It also contains a formal set of axioms defining mapping between different metadata formats for multimedia. The SIOC Core Ontology provides the main concepts and properties required to describe information from on-line communities (e.g., message boards, wikis, weblogs). We use those ontologies together with properties from SIOC, FOAF and Dublin Core to convert into RDF the descriptions of photos, tweets and slides. The link between the media and the event is realized through the `lode:illustrate` property. Figure 3 depicts the description of photos, tweets and slides related to the `ISWC 2011` conference.

3.3. Events Taxonomy

Events are generally categorized in lightweight taxonomies that provide facets when browsing event directories. We manually analyzed the taxonomy used in various sites, namely Facebook, Eventful, Upcoming, LinkedIn[13], Eventbrite[14] and Ticketmaster[15]. Then, we

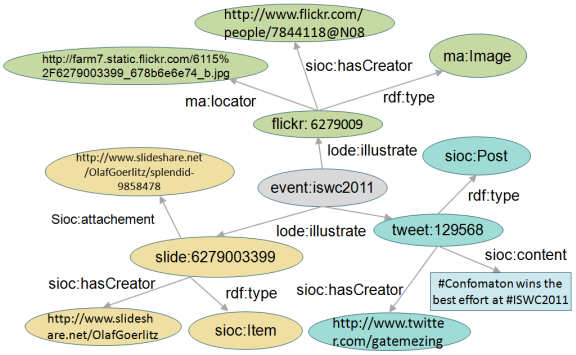

Fig. 3. RDF modeling of photos, tweets and slides associated with the ISWC 2011 Conference

used card sorting techniques in order to build a rich SKOS thesaurus of event categories. This SKOS thesaurus contains axioms expressing mapping relationships with these taxonomies while the terms are defined in our own namespace accessible at (`http://data.linkedevents.org/category`).

4. EventMedia Dataset

EventMedia is a new hub[16] of the Linked Data cloud since September 2010. We use the Last.fm, Eventful, Upcoming and Lanyrd APIs to convert each event description into LODE ontology. We mint new URIs into our own namespace, for example, the URI for events is (`http://data.linkedevents.org/event/`).

Our dataset consists of more than 30 millions RDF triples. All URIs are dereferencable and served as either static RDF files serialized in N3 or as JSON by a RESTful API. The back-end of EventMedia consists of a Virtuoso SPARQL endpoint available at (`http://eventmedia.eurecom.fr/sparql`), a RESTful API available at (`http://eventmedia.eurecom.fr/rest/{resource}`) and powered by the ELDA implementation of the Linked Data API[17]. ELDA provides a configurable way to access RDF data using simple RESTful URLs that are translated into queries to our SPARQL endpoint. The API layer enables associating URIs with processing logic that extracts data from the SPARQL endpoint using one or more SPARQL queries and then serializes the results using the format requested by the client. Figure 4 depicts the architec-

[11] `http://www.w3.org/TR/mediaont-10/`
[12] `http://rdfs.org/sioc/spec/`
[13] `http://www.linkedin.com`
[14] `http://www.eventbrite.com`
[15] `http://www.ticketmaster.com`
[16] `http://ckan.net/package/event-media`
[17] `http://code.google.com/p/linked-data-api`

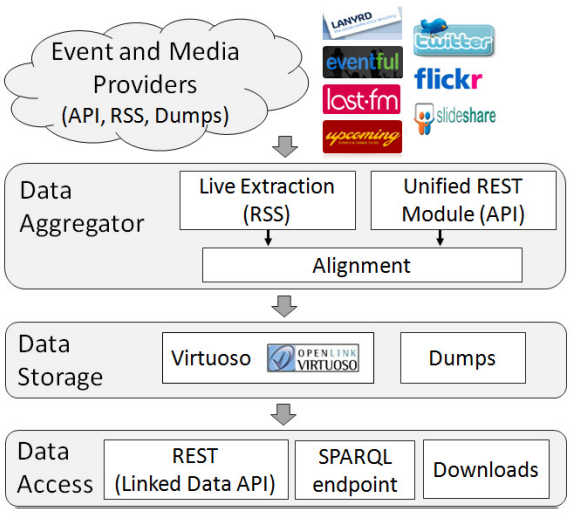

Fig. 4. Overview of the EventMedia components

ture of EventMedia, and Table 1 provides an overview about the number of resources per type and source.

	Event	Agent	Location	Media
Last.fm	57,258	50,150	16,471	1,425,318
Upcoming	13,114	0	7,330	347,959
Eventful	37,647	6,543	14,576	0
Lanyrd	1,167	0	439	537,091
Total	109,186	56,693	38,3816	2,310,368

Table 1
Number of resources per type and source in EventMedia

5. Interlinking

Event directories have overlap in their coverage and it is worthwhile to discover similar events so that one description can complement another. However, discovering similar events from these overlapping but heterogeneous directories imposes some challenges, well-known in instance matching. In addition, we also investigate the enrichment of EventMedia with additional information from open datasets. In our approach, we favour high precision rather than high recall since the cost of missed mapping is lower that the cost of incorrect matching. Statistics about the linksets generated are accessible at (`http://eventmedia.eurecom.fr/dashboard/statistics.html`).

5.1. Interlinking of Event Directories

We create *owl:sameAs* links between events that reflect a high similarity in terms of their factual properties, namely: title, date, location and involved agents. It is worth noting that EventMedia is a challenging dataset due to the presence of a structural heterogeneity (e.g. missing property) and naming variations (e.g. abbreviations, misspellings, different naming conventions). The interlinking was performed using two tools: (i) SILK [5] which draws on a declarative configuration language called Silk-LSL to manually define the linkage rules; (ii) KnoFuss [6] which learns the similarity function based on a semi-supervised genetic algorithm optimizing the precision. We integrated two similarity functions into those tools, namely: a temporal inclusion metric and a string similarity metric described in [7]. The results obtained highlight the time-sensitivity of event reconciliation due to the fact that the time is differently described across multiple websites. Moreover, we note that KnoFuss achieves better performance than SILK thanks to its learning strategy. As a result, the use of KnoFuss on a manually constructed gold-standard of 300 matched events achieves high precision of about 95%, but fair recall of about 75%.

5.2. Enrichment with Linked Data

In order to enrich EventMedia, we perform several interlinking processes using SILK attempting to discover connections between agents and locations with Linked Data. In this context, the key challenge is to resolve the naming conflicts which needs to invoke additional features apart from the instance name. For example, to reconcile the agents, we decide to compare agents' names and descriptions respectively using Jaro and Cosine functions and we set a high threshold to ensure a high precision. Several datasets have been considered such as Musicbrainz, DBpedia, Freebase and Uberblic. Hence, the agent URI which has for label "Radiohead" is interlinked with the DBpedia URI (`http://dbpedia.org/page/Radiohead`) providing information about this band such as its complete discography. Similarly, the datasets being selected to enrich the locations are: DBpedia, Foursquare and Geonames hosting a large amount of geographical information. In the similarity function, we combine the geographical distance and Jaro function applied on labels.

6. Event-Based Applications

The EventMedia dataset has been employed in some web applications designed to enable efficient brows-

ing of an event-based space [8,9,10]. For instance, the EventMedia application [9] delivers different event-centric views (what, where, when and who) and allows users to relive experiences based on media. In fact, people wish to discover events either through invitations and recommendations, or by filtering available events according to their interests [2]. Therefore, the interface allows constraining different event properties (e.g. time, place, category) using, for example, a timeline slider control input and a map grouping markers. Once an event selected, media are presented to convey the event experience, along with social information to provide better decision support. The application is available online at `http://eventmedia.eurecom.fr`. Another application called *Confomaton* follows the same perspective with a focus on conference events. Its goal is to provide a visual summary of a scientific conference including microposts, presentation slides, photos and videos, so that the attendees can catch up with what they could have missed. *Confomaton* is available online at `http://eventmedia.eurecom.fr/confomaton`.

7. Conclusion and Future Work

The integration of event-centric information from social services using Semantic Web technologies has given rise to EventMedia, an open dataset continuously synchronized with recent updates. Several improvements could potentially enhance its quality and usability. Indeed, further vocabularies could be incorporated such as the Ticket Ontology to add meaningful relationships between events and related tickets, or the Allen's vocabulary to express the temporal relationships between events in fine-grained level. Another improvement is to enrich EventMedia using other services such as Youtube, Google+ or Facebook, so that we increase the dataset coverage and more connections could straightforwardly be explored. Finally, we aim to develop a live interlinking framework that aligns in real-time every incoming stream of events with Linked Data.

References

[1] R. Shaw, R. Troncy, and L. Hardman, "LODE: linking open descriptions of events," in *Proceedings of the Semantic Web: 4^{th} Asian Conference, ASWC 2009, Shanghai, China, December 6-9, 2009* (A. Gmez-Prez, Y. Yu, and Y. Ding, eds.), vol. 5926 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), pp. 153–167, Springer Verlag, 2009.

[2] A. Fialho, R. Troncy, L. Hardman, C. Saathoff, and A. Scherp, "What's on this evening? designing user support for event-based annotation and exploration of media," in *Proceedings of the Workshop on Recognising and Tracking Events on the Web and in Real Life, located at the The 6^{th} Hellenic Conference on Artificial Intelligence SETN 2010, Athens, Greece, May 04, 2010* (T. Winkler, A. Artikis, Y. Kompatsiaris, and P. Mylonas, eds.), vol. 624, (Aachen, Germany), pp. 40–54, CEUR Workshop Proceedings, 2010.

[3] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, "The Music Ontology," in *Proceedings of the 8^{th} International Conference on Music Information Retrieval* (S. Dixon, D. Bainbridge, and R. Typke, eds.), (Vienna, Austria), pp. 417–422, Österreichische Computer Gesellschaft, 2007.

[4] M. Doerr, "The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata," in *AI Magazine - Special issue on Ontology Research* (C. Welty, ed.), vol. 24, pp. 75–92, Palo Alto, CA, USA: Association for the Advancement of Artificial Intelligence, 2003.

[5] A. Jentzsch, R. Isele, and C. Bizer, "Silk - Generating RDF Links while publishing or consuming Linked Data," in *Proceedings of the ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, Shanghai, China, November 9, 2010* (A. Polleres and H. Chen, eds.), vol. 658, (Aachen, Germany), pp. 53–56, CEUR Workshop Proceedings, 2010.

[6] A. Nikolov, V. Uren, E. Motta, and A. D. Roeck, "Handling instance coreferencing in the KnoFuss architecture," in *Proceedings of the 1^{st} IRSW2008 International Workshop on Identity and Reference on the Semantic Web, Tenerife, Spain, June 2, 2008* (P. Bouquet, H. Halpin, H. Stoermer, and G. Tummarello, eds.), vol. 422, (Aachen, Germany), pp. 53–56, CEUR Workshop Proceedings, 2008.

[7] H. Khrouf and R. Troncy, "EventMedia Live: reconciliating Events Descriptions in the Web of Data," in *Proceedings of the 6^{th} International Workshop on Ontology Matching (OM-2011) In conjunction with the International Semantic Web Conference (ISWC2011), Bonn, Germany, October 24, 2011* (P. Shvaiko, J. Euzenat, T. Heath, C. Quix, M. Mao, and I. Cruz, eds.), vol. 814, (Aachen, Germany), pp. 250–251, CEUR Workshop Proceedings, 2011.

[8] S. Buschbeck, A. Jameson, R. Troncy, H. Khrouf, O. Suominen, and A. Spirescu, "A demonstrator for parallel faceted browsing." Available online at `http://imash.leeds.ac.uk/event/pdf/Buschbeck_1.pdf`. Presented at the International Workshop on Intelligent Exploration of Semantic Data (IESD, in conjunction with EKAW 2012.

[9] H. Khrouf and R. Troncy, "EventMedia: visualizing events and associated media." Available online at `http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/PostersDemos/iswc11pd_submission_78.pdf`. Presented at Posters & Demonstrations Track of the 10^{th} International Semantic Web Conference (ISWC'11), Bonn, Germany, October 25, 2011.

[10] H. Khrouf, G. Atemezing, G. Rizzo, R. Troncy, and T. Steiner, "Aggregating social media for enhancing conference experience," in *AAAI Technical Report WS-12-02 on Real-Time Analysis and Mining of Social Stream* (A. Zubiaga, D. Spina, M. de Rijke, M. Strohmaier, and M. Naaman, eds.), (Palo Alto, CA, USA), pp. 34–37, Association for the Advancement of Artificial Intelligence, 2012.